Browsing for information on the web and in the file system

Ethan Seifert, Simone Stumpf, Jonathan Herlocker
EECS, Oregon State University
Corvallis, OR
{seiferet, stumpf, herlock}@eecs.oregonstate.edu

Eleanor Wynn
Intel
Hillsboro, OR
eleanor.wynn@intel.com

ABSTRACT
Browsing is one of the methods used for finding and refinding information on the web or in the file local system and there are opportunities to avoid this, particularly if that information is revisited frequently. We present empirical results from a field study contrasting patterns of browsing to local and web information and we qualify the cost that this navigation method incurs. In addition, we provide an improved method for defining revisit behavior and report on the level of revisits during our study. Our findings have implications for solution development that reduce user effort for finding and refinding information.

Author Keywords
Browsing, finding, refinding, navigation, files, web pages.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. H3.3 Information Storage and Retrieval: Information Search and Retrieval

INTRODUCTION
Browsing, alongside with searching, is one of the routine and reoccurring methods used when computer users look for information, either locally in files or on the internet in web pages. Searching for information, especially with current improved desktop searching technology, is often seen as the silver bullet. However, searching does not eliminate all problems. First, users still need to generate appropriate search terms and some users may prefer browsing over searching because it allows information to be found in an iterative and guided way [2, 11]. Second, while most search tools provide help in finding information, they do not provide support when people aim to re-find information. Searching therefore can cut some but not all costs.

In order to access information, much time and effort may be spent on having to first find and then often re-find information yet there may be some opportunities to reduce time and effort for browsing, particularly if that information is revisited frequently. In order to shape better solutions, we are interested in understanding how knowledge workers get to information on the web and local file systems.

In this report, we describe a field study with knowledge workers during which data was logged on their client computers as they browsed for information in Internet Explorer and Windows Explorer. We present results from this study comparing navigation in the web to local navigation and describe differing navigation behavior patterns. We then present an improved method for defining and calculating revisits and report results of revisit levels for our field study. Our results provide implications in terms of potential solution development that aims reduce user effort for finding and refinding information.

RELATED WORK
Our work builds on the findings of previous researchers who have investigated how people access information on the web and local file system.

Information finding on the web has studied the navigation behavior of users employing both qualitative and quantitative approaches [3,11]. Quantitative studies have addressed the relationship between number of web pages visited, the speed of browsing and the number of Internet Explorer windows open [7]. Studies of information refinding have given an account of web visits and revisit behavior by comparing the ratio of new pages vs. previous visited pages [13,5].

For information stored locally, numerous studies have found that users often rely on the classification of information into folders to facilitate retrieval [8], and prefer manual browsing over logical search [2,11]. One reason for this may be that users prefer to navigate to a desired file in small steps using context as a guide [14]. Surprisingly, comparable studies to information finding on the web have rarely been conducted to investigate how information that is stored in local file systems is found and refound.

STUDY SET-UP AND METHODOLOGY
Six knowledge workers, employed in a major high-tech company in a variety of professions such as managers, software engineers and administration, participated in a study over three months where we collected over 2,095 hours of data. Data logs were gathered using the TaskTracer system [6].

TaskTracer keeps track of files and web pages by listening in on most user interaction events in the Microsoft Windows environment and it also logs high-level events in all

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
23 FEB 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Browsing for information on the web and in the file system	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Oregon State University,School of Electrical Engineering and Computer Science,1148 Kelley Engineering Center,Corvallis,OR,97331-5501**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Browsing is one of the methods used for finding and refinding information on the web or in the file local system and there are opportunities to avoid this, particularly if that information is revisited frequently. We present empirical results from a field study contrasting patterns of browsing to local and web information and we qualify the cost that this navigation method incurs. In addition, we provide an improved method for defining revisit behavior and report on the level of revisits during our study. Our findings have implications for solution development that reduce user effort for finding and refinding information.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**5**	

major Microsoft applications. Each event includes data on the time the event occurs, from which component the event originates, and details of the information resource.

Our analysis focused on contrasting specific browsing behaviors. We investigated behavior in Internet Explorer to understand costs in finding information on the web, and browsing through Windows Explorer as an analogue to finding information in the local file system. We have excluded other possible information finding and browsing behavior, such as looking for email messages, from our investigation due to problems with Outlook data logging.

Navigation paths to information

The TaskTracer data logs show users' paths through the web or in the file system. Each time a user browses toward a resource a navigation event is generated. Navigation events occur in Windows Explorer each time the user accesses a folder in the file system hierarchy. In Internet Explorer they take place each time a web page is opened. When a user switches from one window to another afterwards, TaskTracer logs windows focus change events.

The series of navigation events provide the possibility of reconstructing a browsing path. The *navigation path* represents a series of navigations toward an information resource. For each navigation path we can measure its duration and its length (i.e. number of navigation segments) as an indicator of cost.

We define any navigation path as starting with a navigation event and continuing via any number of respective navigation events, terminated by a windows focus change (Figure 1). The focus change is taken as the end of a navigation path because either a) the user has found the information and is opening the resource or b) they have abandoned trying to find the information either temporarily or permanently.

Web navigation paths need an additional stopping condition to reconstruct browsing behavior. Because users can re-use the same Internet Explorer window for consecutive non-related navigation paths, a window focus event may not be generated and therefore several navigation paths may tacked onto each other. Thus, navigation paths in Internet Explorer are also terminated by viewing a web page for longer than 10 seconds, indicating that the user has found and is focusing on some information. (This also prevents transitory pages, used only for navigation purposes, to be terminating pages [10].)

RESULTS

We concentrate on differences in finding and refinding behavior for web pages and for local files. First, we show how often participants navigated to information and their cost of doing so. Second, we qualify the cost of information finding. Then, we report on refinding information by analyzing how often the same information resources were revisited.

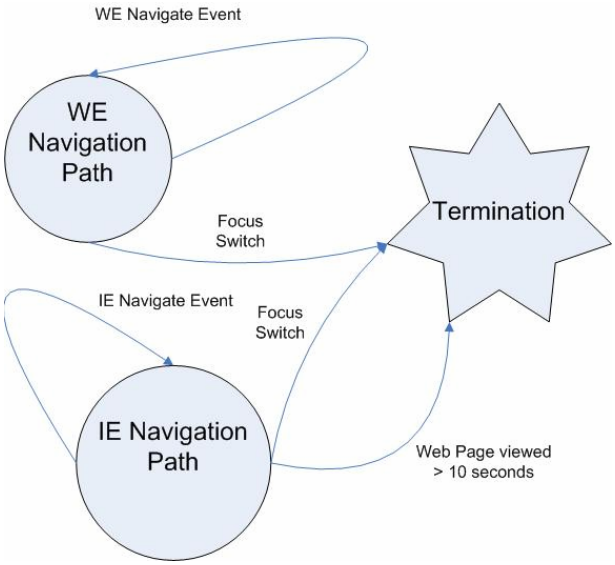

Figure 1. Definitions of Navigation Paths

Browsing to Find Information

Participants conducted an average 215 navigation paths in Internet Explorer over the course of our study, each lasting 6.58 seconds on average (Table 1). The average length of navigation paths was 2 segments, which is in line with previous findings from other researchers [13]. Table 2 shows the results for Windows Explorer navigations for the participants. On average, participants conducted 150 naviga-

Participant	IE Paths	IE Time (seconds)	IE St.Dev. Time	IE Path Length	IE St.Dev. Path Length
P1	14	5.33	4.10	2.25	0.45
P2	499	6.13	4.61	2.62	1.01
P3	121	8.10	5.32	2.68	0.84
P4	273	5.77	4.30	2.41	1.10
P5	266	7.06	3.93	2.43	1.17
P6	122	7.09	3.95	2.68	2.18

Table 1. Navigation in Internet Explorer.

Participant	WE Paths	WE Time (seconds)	WE St.Dev. Time	WE Path Length	WE St.Dev. Path Length
P1	68	8.78	7.88	1.60	2.47
P2	656	11.09	11.14	0.86	1.50
P3	71	12.85	10.64	0.61	1.50
P4	78	16.05	15.03	1.35	1.37
P5	28	15.63	12.68	1.81	1.66
P6	4	8.25	8.54	0.00	0.00

Table 2. Navigation in Windows Explorer.

tion paths, spent 12.11 seconds navigating to information in Windows Explorer, with a path length of 1 segment. A path length of 0 indicates that a folder was opened and they did not have to navigate further to get to the information location. For example, this occurred if a participant opened "My Documents" from the desktop and the file was there.

Discussion

The data shows some interesting user behavior patterns that differed to a great extent between the tools used. One of the major differences is that the time spent browsing in Windows Explorer is on average twice as long as the duration of navigation paths in Internet Explorer, despite the fact that the path length in Windows Explorer is only half as long as in Internet Explorer.

There are various explanations, and potential solutions, for this substantial difference in cost. First, the higher duration in Windows Explorer could be explained by a lack of *information organization* that helped the participants find files again. One popular way of organization is to create a folder hierarchy. In our case, low path lengths could indicate a flatter file system resulting in shorter path lengths (i.e. the participants didn't have to click through several folders to the target file in a flat hierarchy). In turn a flatter file system may also suggest that more files have to be browsed at one location, increasing time to find them. Aiding the user in creating appropriate organizational mechanisms (e.g. through suggesting organization based on activities) may alleviate this problem.

However, this does not fully explain why it should take considerably longer to find information in the local file systems which are organized by the users themselves, in contrast to web information which is organized by someone else (and hence may take longer to make sense of). This leads us to suspect that this differing pattern may be due to *information duplication.* When looking for information on the web there are often several pages that could provide the necessary information. Finding information within a local file system, however, means that usually only one particular information resource will be the right one. Thus, we have a different termination condition for information finding on the web and on the local file system, and if many pieces of the same information are available, the time spent browsing to any one of them decreases.

In addition, there is a lack of *information scent* [4] for local files that enables the user to determine whether the are on the right track. Typically, finding information in local files is based on making judgments given short, and sometimes ambiguous, folder and file names, whereas finding information on the web involves looking through rich media that provides more information to lead one into the right direction. Since information scent is absent in the local file system, the user may take longer to browse in Windows Explorer.

The Cost of Navigating

We analyzed how much time participants spent navigating in Internet Explorer and Windows Explorer in relation to their overall computer usage. (We excluded times when there were no keypresses or mouse clicks during a 15 minute time interval from our calculations.) We found that the average rate of time spent on navigation is 0.9% in relation to their overall time spent on the computer. Individual rates vary considerably between participants but it is usually not higher than 2%.

Discussion

This low figure for navigation costs was initially surprising to us. It shows that actual navigation costs may bear relatively little resemblance to the high level of time and effort that our participants perceived them to be.

This may be due to a number of reasons. First, these costs are likely to be an underestimate. Data from navigation in email clients, other browsers or applications has not been included in our analysis. Furthermore, it has been noted that the amount of scrolling within information can be very high [9] but this is not included in our navigation costs. Second, the perceived time may be felt to be much higher since it is unproductive time – browsing is a means to an end, not a goal in itself. Lastly, if something goes wrong in browsing, then it may influence the overall perception of cost. For example, one of the worst cases in Windows Explorer in our study took 156 seconds and had a navigation path over 23 segments long!

Browsing to Find Information Again

A way of reducing costs to knowledge workers is to find ways to cut down on the amount of browsing they have to do. For example, costs could be reduced by providing mechanisms, such as bookmarks, history lists, etc., that shorten the path length for revisited information.

One established way of calculating revisits to web information includes each page if it has been viewed once before [13]. Following this approach a revisit rate is determined by taking the ratio of all revisited pages over the number of total pages visited. We believe that this way of estimating revisits results in an inflated figure since it makes no distinction about the usefulness of the pages: it includes in the count "visits" to the default homepage and also transitory pages that are only used for navigation purposes.

In order to give a better estimate of revisits we developed a different method. We only count revisits to the destination of the navigation path – for example, we do not include revisits to transitory stages leading up to the destination, nor do we match on similar but not identical destinations. Analyzing the destination of a navigation path allows for a more precise measure by ignoring web pages or folders common to several paths. From this we can calculate a conservative revisit rate, which is the percentage of revisited destinations over all visited destinations. We also provide a frequency of how often this information is revisited. Our measure is the

Participant	IE Re-visits	IE Revisit rate	Average frequency of revisit
P1	3	21%	1.00
P2	247	49%	4.57
P3	24	20%	1.85
P4	120	44%	4.00
P5	59	22%	2.19
P6	50	41%	2.50

Table 3. Revisits in Internet Explorer.

Participant	WE Revisits	WE Revisit rate	Average frequency of revisit
P1	31	46%	2.38
P2	321	49%	3.15
P3	25	35%	1.67
P4	49	63%	3.50
P5	8	29%	1.60
P6	1	25%	1.00

Table 4. Revisits in Windows Explorer.

minimum that could be saved by revisits to the destination.

So how much of these costs could potentially be saved? In order to answer this question we analyzed revisit behavior within Internet Explorer (Table 3) and Windows Explorer (Table 4) during the field study using our approach. Revisits to the same destination within Internet Explorer amounted to 33% on average, and the same page was revisited 3 times on average. We further found that on average 41% of destination folders are revisited twice on average.

Discussion

Our results on revisits in Internet Explorer differ considerably from revisit results reported previously [13,5], which claimed that 58% or even over 80% of visits are revisits. Since we do not include transitory pages in our calculation, this suggests that at least a quarter of costs – and possibly even higher levels – in revisits are attributable to navigation overheads and this could be reduced substantially. In addition, our quantitative findings on revisit behavior to information on the local file system show that considerable savings in time and effort could be made. Since we know where the information is located, potentially most of navigation to local information is an overhead cost.

Exploiting machine learning may be a fruitful avenue to cut these kinds of costs. Research within the TaskTracer project has already investigated the feasibility of some solutions for cutting down on refinding information on the web and in the local file system [1,10].

CONCLUSIONS

Our field study provides a number of interesting results. We found that participants' navigation path lengths were shorter in Windows Explorer than in Internet Explorer, yet it took more time to get to the information. This suggests that browsing behavior for finding and refinding information needs to be supported in ways that take these differences into account.

Through our improved method of calculating revisits, we have found that some navigations costs appear to be unavoidable. A large number of web pages were never visited again. This is where search could provide most impact, by reducing the navigation cost to new information.

However, costs could be substantially reduced through making use of revisits and any improvements on these costs alone could have substantial impact on users' perceptions.

ACKNOWLEDGMENTS

We thank the participants of our study. We also thank Julie Lynn for her help. This work was supported by Intel, by NSF IIS-0133994, and by DARPA, HR0011-04-1-0005, NBCHD030010.

REFERENCES

1. Bao, X., Herlocker, J.L., Dietterich, T.G. Fewer Clicks and Less Frustration: Reducing the Cost of Reaching the Right Folder. *Proc. IUI*, (2006), 178-185.
2. Barreau, D., Nardi, B. A. Finding and reminding: file organization from the desktop. *SIGCHI Bull., 27(3)*, (1995), 39-43.
3. Bruce, H., Jones, W., Dumais, S. Information behavior that keeps found things found. *Information Research, 10(1)*, (2004).
4. Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. 2001. Using information scent to model user information needs and actions and the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, United States). CHI '01. ACM Press, New York, NY, 490-497.
5. Cockburn, A., Greenberg, S., Jones, S., McKenzie, B., Moyle, M. Improving Web Page Revisitation: Analysis, Design & Evaluation. *IT&Society, 1(3)*, (2003), 159-183.
6. Dragunov, A.N., Dietterich, T.G., Johnsrude, K., McLaughlin, M., Li, L., Herlocker, J.L. TaskTracer: A desktop environment to support multi-tasking knowledge workers. *Proc. IUI*, ACM Press (2005), 75–82.
7. Hawkey, K. , Inkpen, K. Web browsing today: the impact of changing contexts on user activity, *Ext. Abstracts CHI*, (2005).
8. Jones, W., Phuwanartnurak, A. J., Gill, R., Bruce, H. Don't take my folders away!: organizing personal information to get things done. *Ext. Abstracts CHI*, (2005).
9. Ko, A.J., Aung, H.H., Myers, B.A. Eliciting Design Requirements for Maintenance-Oriented IDEs: A Detailed Study of Corrective and Perfective Maintenance Tasks. *Proc. ICSE*, (2005).
10. Lettkeman, A.T., Stumpf, S., Irvine, J., Herlocker, J.L.

Predicting Task-Specific Webpages for Revisiting. *Proc. AAAI*, (2006).

11. Ravasio, P., Schar, S., Krueger, H. In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Trans. Computer-Human Interaction, 11(2)*, ACM Press (2004), 156–180.
12. Sellen, A. J., Murphy, R., Shaw, K. L. How knowledge workers use the web. *Proc. CHI* (2002), 227-234.
13. Tauscher, L., Greenberg S. (1997) How people revisit web pages: empirical findings and implications for the design of history systems. *Int. Journal Human-Computer Studies, 47*, 97-137.
14. Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R., The perfect search engine is not enough: a study of orienteering behavior in directed search. *Proc. CHI*, ACM Press (2004), 415–422.